HIGH SCHOOL EDITION
Journal of Student Research
Volume 11 Issue 4 (2022)

Using a Logistic Regression and K Nearest Neighbor Model to Accurately Diagnose Breast Cancer

Shridhula Srinivasan[1], Mridula Srinivasan[1] and Govind Tatachari[#]

[1] Monta Vista High School, Cupertino, CA, USA
[#]Advisor

ABSTRACT

Breast cancer is one of the most dangerous and rapidly growing diseases in the world. Diagnosing breast cancer is expensive, difficult, and time-consuming. However, artificial intelligence and machine learning algorithms can help physicians to diagnose people with breast cancer at an early stage which will help people to avoid exhaustive treatments. The objective of our research was to classify if someone has malignant or benign cancer. We used the Wisconsin Breast Cancer dataset which was obtained from the UCI repository to create models using supervised learning. We used K Nearest Neighbors, and Logistic Regression algorithms to obtain a model with high accuracy. Both the models had an accuracy of 97%. In the future, the model can be enhanced to be more accurate and accessible to people. This research can help others to create models to predict various other cancers. In the future, we would also like to improve the model by using other methods like image recognition and reducing the input from the user to make it more accurate and accessible.

Introduction

Annually more than 255,000 women and 2,300 men are diagnosed with breast cancer. It is crucial to diagnose breast cancer ahead of time because it increases the chances of survival. Moreover, it increases the treatment options available. Furthermore, early diagnosis will allow people to avoid exhausting treatments. The current treatments for breast cancer are surgery, chemotherapy, hormonal therapy, and radiation therapy. In hormonal therapy, the cancerous cells are blocked from getting hormones to expand and grow, and in radiation therapy x-rays are used to kill the cancerous cells. These treatments are expensive, tiring, and painful. Using machine learning and artificial intelligence, physicians can diagnose breast cancer more accurately and easily. The models can also increase accessibility by allowing physicians to diagnose breast cancer with minimal equipment.

Our main objective was to try multiple different models on our dataset to get the most accurate and precise model that helps patients to be diagnosed with breast cancer before it gets out of control. This will help people to avoid exhausting treatments and instead get access to treatments as soon as possible. In our experiment, we used the Breast Cancer Wisconsin dataset from the UCI repository. The data set includes 569 instances. The input values in the dataset are all computed from an image of a fine needle aspiration test performed on the patient. A fine needle aspirate is a procedure used to draw cells and fluid from under the skin. The values are descriptions of the characteristics of the cells drawn from the patient's skin. The attributes in the data set include the radius, texture, perimeter, area, smoothness, compactness, concavity, number of concave portions, symmetry, and the fractal dimension of the cell's nucleus.

In our project, we used supervised learning and classification algorithms. Supervised learning consists of classification and regression algorithms. Regression algorithms are used to predict values with a continuous

range of outcomes, while classification algorithms are used when the outcome is discrete and consists of categories. In our dataset, our model has to predict if a person has a malignant or benign cancer. In our research we used a logistic regression model, and a K nearest algorithm to find the most precise model which could accurately predict whether a patient has breast cancer or not. Both of our models performed relatively well, with an accuracy of around 97%. In the future, we would like to try using various other datasets and models to predict breast cancer. We would also like to predict not only the type of breast cancer the patient has but also the severity of their condition through a range of values. In addition, to make our model accessible, images and other factors can be used to predict the patient's condition.

Methods

Our objective was to create a model to predict if a patient has malignant or benign cancer. For our research, we made use of the Breast Cancer Wisconsin dataset from the UCI repository. We first cleaned the data by removing columns with no values. We also changed the prediction to numerical values, with malignant prediction set to 1 and benign cancer to 0. We made use of StandardScaler in sklearn to resize the distribution of the input values. The standard scaler will make the standard deviation 1 and remove the mean from the data set. Next, we performed data analysis to gain a deeper understanding of which values played a huge role in determining if the cancer was malignant or benign (See Figure 1). We used seaborn, matplotlib, and numPy to conduct our data analysis. From our data analysis, we found that the mean of all input values was greater for malignant cancer except for the fractal dimension of the nucleus. We also looked at the mean, minimum, maximum, and standard deviation to find trends in the data set. The mean of the radius, perimeter, area, concave points, and compactness was much greater for malignant cells (See Figure 2).

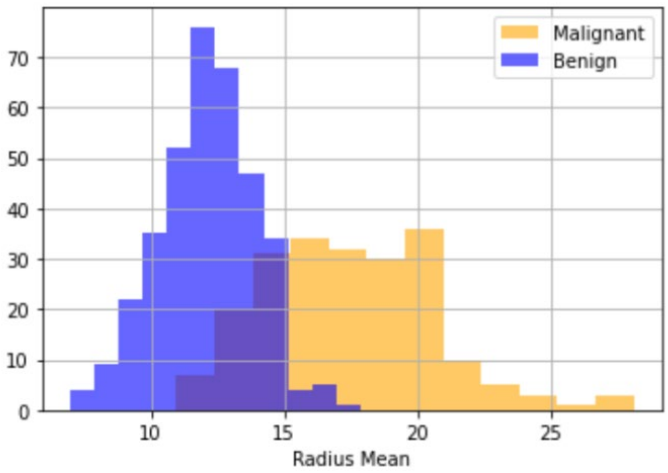

Figure 1. The mean radius of patients with a malignant tumor and benign tumor. Patients with a benign tumor tend to have a radius mean lower than patients with a malignant tumor.

We then split our data into training and testing sets. We did an 80:20 split for the training to the testing set. Different classification algorithms were used to get the model with high accuracy. The logistic regression predicts the binary outcome by using independent input values. The logistic regression algorithm reports the probability of the event and helps to identify the independent variables that affect the dependent variable the most. The K Nearest Neighbors predicts the outcome by calculating the distance from the testing values to the

training set and selecting the values closer to the testing value. An error plot was used to identify the k value that increases the accuracy while making sure that the model is not overfit for the data. For each algorithm, a confusion matrix and a classification report were created to analyze the outcome and to determine how accurate the model is. The log loss function from sklearn was also used to evaluate the logistic regression model.

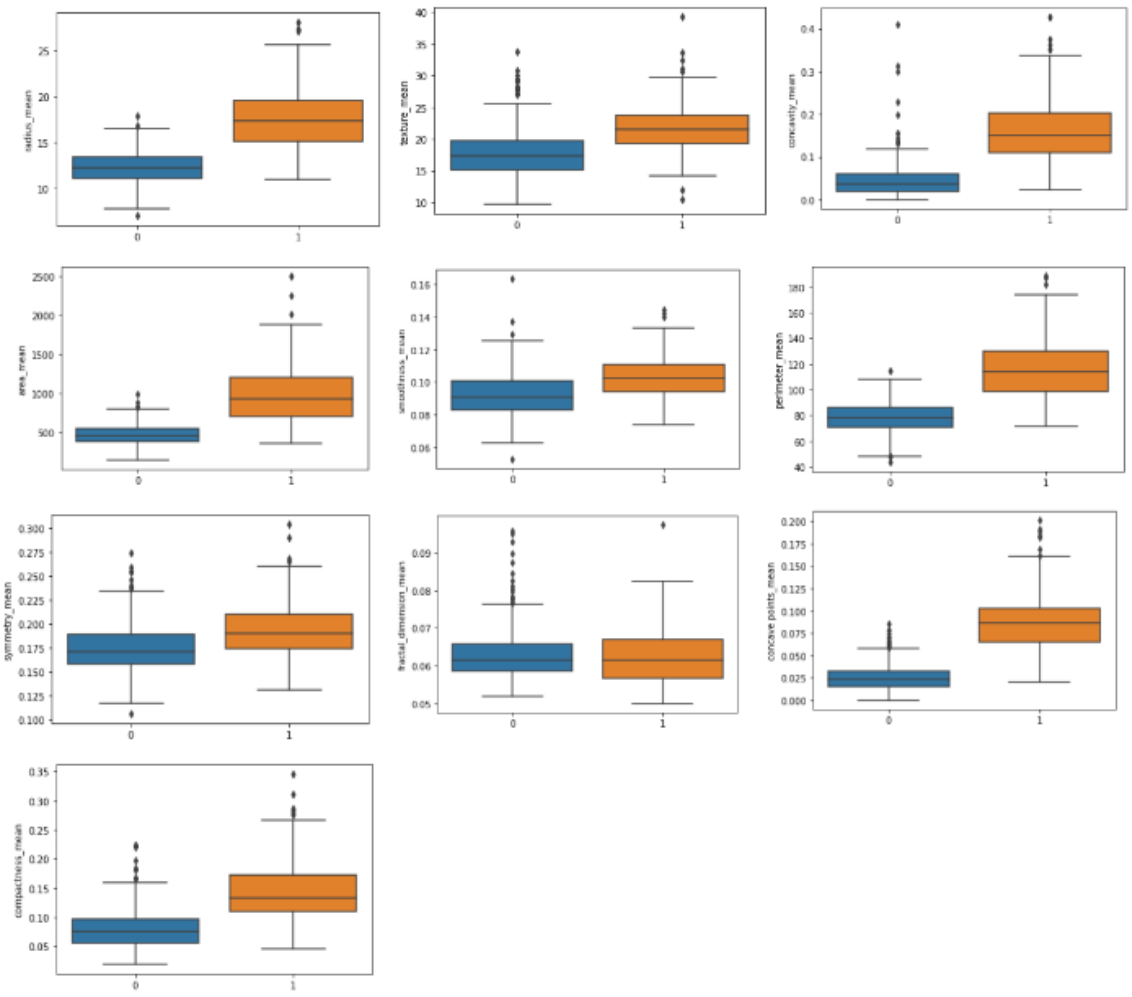

Figure 2. Data exploration: All attributes for malignant and benign patients were plotted side by side to find trends in the data.

Discussion

We split our data set into two groups, one for training and the other for testing our model. Having drastically more training data causes overfitting where the model picks up on all the minor details and has difficulty when new data is presented. On the other hand, having too little data for the training set can cause underfitting, which can not create an accurate model as the model is not able to find a correlation between the input and output. For our model, we did an 80:20 split of the training data to the testing data. This ensures that the model has high accuracy and can better predict the new input.

The dataset's quality could have affected the model used to predict the output. A clean dataset with more input values could help improve the model and accuracy. The logistic regression model, and the K nearest algorithm had an accuracy of 0.97. Our research will help patients to prevent their conditions from worsening and also gain access to effective treatment options to improve their symptoms. This can also help others to use datasets to create algorithms and devices to help predict diseases in a fast and efficient manner.

In the future, we would like to improve our model by using various other datasets that have different input factors to predict the output. We could also use datasets that will help to diagnose patients in a spectrum of severity based on their conditions instead of two discrete outcomes. This will help people to take preventative medicines to stop the progression or development of cancer. In addition, we would like our model to recognize images and other inputs to make it more accessible to more patients around the world. The results from our research can help physicians to easily diagnose patients with cancer. The results can also help researchers to create other models and new algorithms in the future to diagnose different types of cancer.

Results

For our machine learning model, we used a dataset from the Breast Cancer Wisconsin dataset from the UCI repository. In our dataset, we had 10 input values that the model took in as input to predict the expected output of either a yes or no. It predicts yes if they have malignant cancer and no otherwise. The dataset had an 80:20 split for training and testing the data. The data was trained using various different models. We trained it using a logistic regression model, and the K nearest algorithm.

The Classification report visualizer reports four values, which include precision, recall, f1-score, and support. The precision measures the ratio of true positives to the sum of true and false positives in the dataset. The recall score is the ratio of true positives over the sum of the true positives and false negatives. The f1-score is the average of precision and recall. The support score is the number of samples of true responses that are there in the dataset. The weighted recall score, f1-score, and precision score for the logistic regression is 0.97. The weighted average support score was 171. The weighted recall score, f1-score and precision score for the K nearest algorithm is 0.97. The weighted average support score was 171 (See Figure 4).

	precision	recall	f1-score	support
0	0.98	0.97	0.98	108
1	0.95	0.97	0.96	63
accuracy			0.97	171
macro avg	0.97	0.97	0.97	171
weighted avg	0.97	0.97	0.97	171

	precision	recall	f1-score	support
0	0.98	0.97	0.98	108
1	0.95	0.97	0.96	63
accuracy			0.97	171
macro avg	0.97	0.97	0.97	171
weighted avg	0.97	0.97	0.97	171

Figure 4. Classification Report Calculation for Logistic Regression Algorithm and K Nearest Algorithm

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

UCI Machine Learning Repository: Breast Cancer wisconsin (diagnostic) data set. (n.d.). Retrieved June 25, 2022, from https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

Supervised machine learning - javatpoint. www.javatpoint.com. (n.d.). Retrieved June 25, 2022, from https://www.javatpoint.com/supervised-machine-learning

Breast cancer early detection and diagnosis: How to detect breast cancer. American Cancer Society. (n.d.). Retrieved June 25, 2022, from https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection.html

Logistic regression. Logistic Regression - an overview | ScienceDirect Topics. (n.d.). Retrieved June 25, 2022, from https://www.sciencedirect.com/topics/computer-science/logistic-regression#:~:text=Logistic%20regression%20is%20a%20process,%2Fno%2C%20and%20so%20on.

Statistical Data Analysis Techniques in machine learning. Analytics Vidhya. (2021, June 24). Retrieved June 25, 2022, from https://www.analyticsvidhya.com/blog/2021/06/must-know-statistical-data-analysis-techniques-in-machine-learning/

Mayo Foundation for Medical Education and Research. (2022, April 27). *Breast cancer*. Mayo Clinic. Retrieved June 30, 2022, from https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475#:~:text=Most%20women%20undergo%20surgery%20for,before%20surgery%20in%20certain%20situations.